# Project #428 – AudiComm

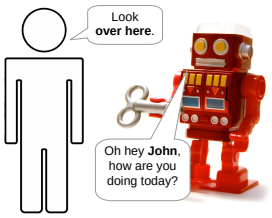
## Audio for Communication and Environment Perception

**K. Diepold, M. Durkovic[1], A. Knoll, M. Giuliani[2], F. Wallhoff, J. Geiger[3]**

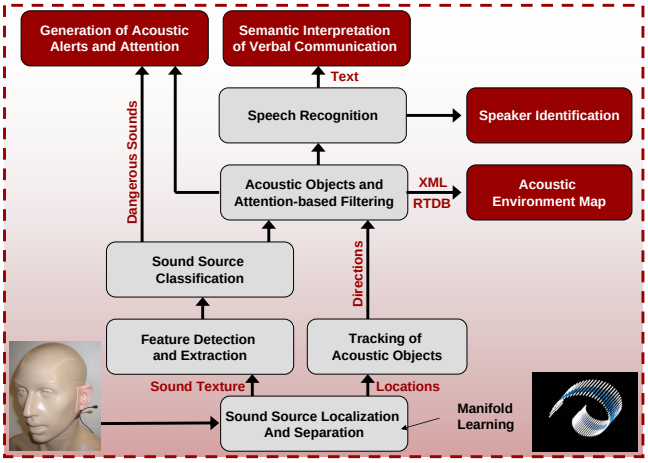
{kldi,durkovic}@tum.de[1], {knoll,giuliani}@in.tum.de[2], {wallhoff,geiger}@tum.de[2]

## Research Goals

- **Sound localization**
  with head-related transfer functions (HRTF)
- **HRTF Database**
- **Speaker identification**
  with Gaussian mixture models
  online learning of new speakers
- **Sound source classification**
- **Speech processing**
  with combinatory categorial grammar
- **Acoustic environment map**

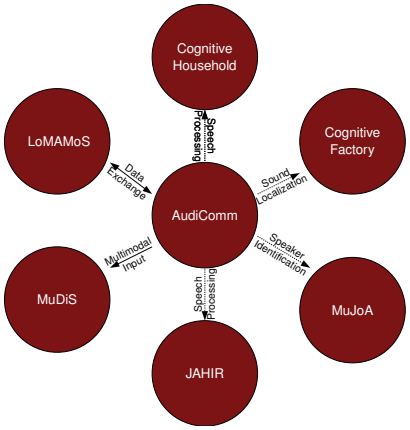



## Real Time Architecture




## Connection to other Projects

- Data exchange with LoMAMoS
- AudiComm provides additional multimodal input for MuDiS
- Speech processing for JAHIR
- In the future we plan to provide
  - Speech processing for Cognitive Household
  - Sound Localization for Cognitive Factory
  - Speaker Identification for MuJoA



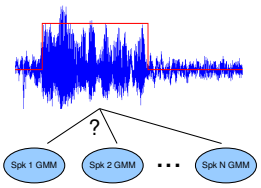

## Contribution to Demonstrators

- AudiComm software is available to the whole cluster, in the future AudiComm can provide for
- **Multi Joint Action**
  - Improvement of human-robot interaction through better acoustic environment perception
  - Advanced dialogue management in collaboration with MuDiS
  - Speaker identification to customize robot behaviour
- **Cognitive Factory**
  - Sound localization for improved security for human worker in JAHIR
  - Speaker identification for customization of assembly plan for worker
  - Specialized grammar-based speech processing for joint assembly tasks
- **Cognitive Household**
  - Software to localize household inhabitants with microphones (for example when it is dark)
  - Speech processing to give robot commands
  - Speaker identification to distinguish between inhabitants and guests

## Methods and Results

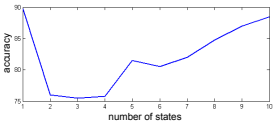
### Sound Localization

- Sound source localization using generalized cross correlation
- Development towards binaural robotic hearing
- Efficient HRTF database personalization for improved binaural sound localization
- Recording of custom HRTFs in noisy environments for online HRTF learning




### Speaker Identification

- Energy-based speech/silence detection
- Features: 12 MFCCs, Energy, $\Delta$, $\Delta\Delta$
- Gaussian mixture models with 4 mixtures
- Tested with different numbers of states for the models (see picture)
- Recognition results on baseline database (YOHO): $\sim$ 90% with 10 speakers
- Realtime processing
- Implemented with HTK




### Speech Processing

- Grammar for spatial language available
- Language grounding with audible cues
- Multimodal processing
- Inclusion of context information

$$\omega_{l1:situation}(\text{look-verb} \wedge \langle \text{MOOD} \rangle \text{ imp} \wedge \langle \text{PATIENT} \rangle x1 : sem - obj \wedge \langle \text{PATIENT2} \rangle ( p1 : animate - being \wedge \text{pron} \wedge \langle \text{NUM} \rangle \text{ sg} \wedge \langle \text{PERS} \rangle \text{ 1st} \wedge \langle \text{PREP} \rangle \text{ to-prep} ))$$

### Demonstration

- Includes sound localization, speaker identification, and speech processing
- Runs on one computer $\Rightarrow$ portable to any computer running linux
- Sound localization customizable $\Rightarrow$ works with any stereo microphone
- Recognition of 4 different speakers + unknown speaker
- Grammar parses sentences that contain spatial description (“look to me”) and grounds them in audio data

## Future Plans

- Collection of data for HRTF database
- Acoustic event detection (for example distinguishing human and non-human noises)
- Development of an acoustic environment map
- Theory of language grounding in acoustic events
- Grammar with advanced spatial language description for CoTeSys demonstrators

Data Processing[1], Robotics and Embedded Systems[2], Human-Machine Communication[3]

Technische Universität München[1,2,3], Munich, Germany